



Document Summary



New
Search



Help

[Preview Claims](#)

[Preview Full Text](#)

[Preview Full Image](#)

Email Link: 

Document ID: J P 04-018673 A2

Title: METHOD AND DEVICE FOR EXTRACTING TEXT INFORMATION

Assignee: HITACHI LTD

Inventor: MORIMOTO YASUTSUGU
YAMANO FUMIYUKI

US Class:

Int'l Class: G06F 15/401 A

Issue Date: 01/22/1992

Filing Date: 05/11/1990

Abstract:

PURPOSE: To easily detect a part corresponding to an abstract by storing information of relations between sentences in a document and abstracted sentences and obtaining sentences of the original document corresponding to abstracted sentences in accordance with this information and outputting text information such as a list of words appearing in these sentences.

CONSTITUTION: A document is retrieved from an English text file and is subjected to morpheme analysis, and the results are stored in an appearing word table. The range of sentences in the original document corresponding to each abstracted sentence is determined and the abstract is displayed, and text information is extracted and displayed. Thus, a part of original sentences is accurately displayed to efficiently acquire text information.

(C)1992,JPO&Japio

⑨ 日本国特許庁(JP)

⑩ 特許出願公開

⑫ 公開特許公報(A)

平4-18673

⑤ Int. Cl.⁵

識別記号

庁内整理番号

④ 公開 平成4年(1992)1月22日

G 06 F 15/401

7056-5L

審査請求 未請求 請求項の数 17 (全 28 頁)

⑥ 発明の名称 テキスト情報抽出方法および装置

⑦ 特 願 平2-122411

⑧ 出 願 平2(1990)5月11日

⑨ 発 明 者 森 本 康 嗣 神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

⑩ 発 明 者 山 野 文 行 神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

⑪ 出 願 人 株式会社日立製作所 東京都千代田区神田駿河台4丁目6番地

⑫ 代 理 人 弁理士 有近 紳志郎

明 細 書

1. 発明の名称

テキスト情報抽出方法および装置

2. 特許請求の範囲

1. 文書と、その文書の抄録と、前記文書中の文章と前記抄録中の文の関係情報とを記憶する記憶ステップと、

前記抄録中の文のうちの少なくとも1つを選択する抄録文選択ステップと、

その選択された抄録文に対応する前記文書中の文章を前記関係情報を用いて前記文書から抽出する対応文章抽出ステップと、

その抽出された対応文章についてのテキスト情報を取り出して出力するテキスト情報出力ステップと

を有することを特徴とするテキスト情報抽出方法。

2. 記憶ステップにおいて、文書中の文と抄録文の1対1の対応情報を記憶し、その記憶した1対1の対応情報から抄録文に対応する文

書中の文を調べてその文を含む複数の文を対応文章の範囲として決定し、その対応文章の範囲を関係情報として記憶する請求項1のテキスト情報抽出方法。

3. 文書の抄録を自動作成し、その抄録の自動作成の過程で得られた情報から前記文書中の文と抄録文の関係情報を自動作成する請求項1または請求項2のテキスト情報抽出方法。

4. 複数の文書の中から1つの文書を選択するための文書選択情報を入力して文書を選択する文書選択ステップをさらに有し、

前記選択した文書の抄録を自動作成し、その抄録の自動作成の過程で得られた情報から前記文書中の文と抄録文の関係情報を自動作成する請求項1または請求項2のテキスト情報抽出方法。

5. ある文書についてのテキスト情報出力ステップ中に、他の文書についての抄録文選択ステップと、対応文章抽出ステップと、テキスト情報出力ステップとを再帰的に実行するべ

- く、複数の文書の中から前記他の文書を選択する他文書選択ステップをさらに有する請求項4のテキスト情報抽出方法。
6. 文書選択情報により選択した文書の抄録を自動作成する際に、前記文書選択情報を利用する請求項4または請求項5のテキスト情報抽出方法。
7. テキスト情報の種類をユーザが選択するテキスト情報選択ステップをさらに有する請求項1から請求項6のいずれかのテキスト情報抽出方法。
8. 抄録文に対応する文書中の文を含む段落の全文を対応文章の範囲として決定する請求項1から請求項7のいずれかのテキスト情報抽出方法。
9. 抄録文に対応する文書中の文より所定の前方相対値だけ前の文から所定の後方相対値だけ後の文までの全文を対応文章の範囲として決定する請求項1から請求項7のいずれかのテキスト情報抽出方法。

— 3 —

15. 文書と、その文書の抄録と、前記文書中の文章と前記抄録中の文の関係情報とを記憶する記憶手段と、

その記憶手段に記憶した抄録を出力する抄録出力手段と、

その抄録出力手段により出力された抄録中の文のうちの少なくとも1つをユーザが選択する抄録文選択手段と、

その選択された抄録文に対応する前記文書中の文章を前記関係情報を用いて前記文書から抽出する対応文章抽出手段と、

その抽出された対応文章についてのテキスト情報を取り出して出力するテキスト情報出力手段と

を具備してなることを特徴とするテキスト情報抽出装置。

16. 記憶手段が、文書中の文と抄録文の1対1の対応情報を記憶し、その記憶した1対1の対応情報から抄録文に対応する文書中の文を調べてその文を含む複数の文を対応文章の範

10. 対応文章の最初の語句または最後の語句が所定の語句であるときは、その所定の語句に応じて対応文章の範囲を前後に拡張し、新たな対応文章とする対応文章拡張ステップをさらに有する請求項1から請求項9のいずれかのテキスト情報抽出方法。
11. 対応文章をユーザに提示し、ユーザからの変更指示入力に応じて対応文章の範囲を前後に変更し、新たな対応文章とする対応文章変更ステップをさらに有する請求項1から請求項10のいずれかのテキスト情報抽出方法。
12. テキスト情報が、対応文章の範囲内の各文である請求項1から請求項11のいずれかのテキスト情報抽出方法。
13. テキスト情報が、対応文章の範囲内に出現する単語のリストである請求項1から請求項11のいずれかのテキスト情報抽出方法。
14. 文書中の各段落の先頭の文を抽出して抄録とする請求項1から請求項13のいずれかのテキスト情報抽出方法。

— 4 —

囲として決定し、その対応文章の範囲を関係情報として記憶する請求項15のテキスト情報抽出装置。

17. テキスト情報出力手段が、選択された抄録文とテキスト情報とを対照可能に同時出力する請求項15または請求項16のテキスト情報抽出装置。

3. 発明の詳細な説明

[産業上の利用分野]

本発明は、テキスト情報抽出方法および装置に関し、さらに詳しくは、ある文書についてのテキスト情報をその文書の抄録を利用して容易に得られるようにしたテキスト情報抽出方法および装置に関する。

[従来の技術]

従来、ある文書における重要箇所とか、その重要箇所に出現する単語のリストとかのテキスト情報を得たい場合、人が文書を解読して重要箇所を取り出したり、その取り出した重要箇所を1つの文書として出現単語リスト自動抽出装置に入力し

— 5 —

— 588 —

— 6 —

て単語リストを得たりしている。

また、従来、抄録を読んでいた元の文書の文章を参照したい場合、人が文書を解読して対応部分を見つけ出している。

〔発明が解決しようとする課題〕

しかし、人が文書を解読して重要箇所を取り出したり、抄録に対応する部分を見つけ出すのは非常に手間がかかる問題点がある。

そこで、本発明の目的は、ある文書における重要箇所とか、その重要箇所に出現する単語のリストとかのテキスト情報を容易に得られるようにすると共に、抄録から元の文書の文章を容易に参照できるようにするテキスト情報抽出方法を提供することにある。また、そのテキスト情報抽出方法を好適に実施するテキスト情報抽出装置を提供することにある。

〔課題を解決するための手段〕

本発明は、文書と、その文書の抄録と、前記文書中の文章と前記抄録中の文の関係情報とを記憶する記憶ステップと、前記抄録中の文のうちの少

なくとも1つを選択する抄録文選択ステップと、その選択された抄録文に対応する前記文書中の文章を前記関係情報を用いて前記文書から抽出する対応文章抽出ステップと、その抽出された対応文章についてのテキスト情報を取り出して出力するテキスト情報出力ステップとを有するテキスト情報抽出方法を提供する。

また、本発明は、文書と、その文書の抄録と、前記文書中の文章と前記抄録中の文の関係情報とを記憶する記憶手段と、その記憶手段に記憶した抄録を出力する抄録出力手段と、その抄録出力手段により出力された抄録中の文のうちの少なくとも1つをユーザが選択する抄録文選択手段と、その選択された抄録文に対応する前記文書中の文章を前記関係情報を用いて前記文書から抽出する対応文章抽出手段と、その抽出された対応文章についてのテキスト情報を取り出して出力するテキスト情報出力手段とを具備してなるテキスト情報抽出装置を提供する。

— 7 —

〔作用〕

本発明のテキスト情報抽出方法および装置では、ある文書中の文とその抄録中の文（抄録文）との関係情報を記憶しておき、抄録文をユーザが選択すると、その抄録文に対応する元の文書の文章を前記関係情報より求め、その元の文書の文章や、その文章に出現する単語のリストなどのテキスト情報を出力する。

そこで、ある文書の重要箇所や、その重要箇所に出現する単語のリストなどを抄録を利用して容易に得られるようになる。

また、抄録文に対応する元の文書の文章を容易に参照できるようになる。

〔実施例〕

以下、図に示す実施例により本発明を詳細に説明する。なお、これにより本発明が限定されるものではない。

第11図は、本発明の第1実施例のテキスト検索・表示システム1000のブロック図である。

図中、1は処理装置、2はディスプレイ装置、

— 8 —

3はキーボード装置、4はメモリ、6は英文テキストファイル、7はストップワードファイル、8は接続語句ファイル、9は辞書ファイルである。

メモリ4は、文書検索プログラム401と、抄録作成プログラム402と、対応範囲決定プログラム403と、テキスト情報抽出プログラム405と、出現単語テーブル406と、単語頻度テーブル407と、文得点テーブル408と、抄録文テーブル409と、対応範囲テーブル410と、表示単語テーブル411と、キーワードテーブル412とを含んでいる。

英文テキストファイル6は、複数の英文の文書60と、各文書毎の管理情報テーブル61とを含んでいる。この管理情報テーブル61は、第12図に示すように、文書の各文についての文番号とその文が属する段落の段落番号とを対応付けたものである。

第1図は、上記テキスト検索・表示システム1000により英文テキストファイル6から文書を検索し、抄録を作成し、その抄録を表示すると共

— 9 —

— 589 —

— 10 —

に、その抄録に対応する元の文章を英文テキストファイル6から抽出して表示する処理のフロー図である。

ステップ11では、英文テキストファイル6から文書を検索する。検索方法は、テキストデータベース検索システム等で公知の検索方法（例えば、検索キーを入力し、その検索キーと予め文書に付与しておいた文書キーとを比較して、一致した文書を取り出す方法等）を利用できる。

ステップ12では、前記取り出した文書の各文を形態素解析し、各文の構成単語を出現単語テーブル406に出現順に格納する。すなわち、第13図に示すように、文書60の各文の単語を切り出して順に出現単語テーブル406に格納する。形態素解析の方法は、例えば特開昭58-40684号公報等に開示の方法を利用できる。

ステップ13では、前記文書の抄録を作成する。

第2図は、抄録作成処理の一例である。

まず、ステップ41では、文書の文中に出現する単語とその出現頻度を求め、第14図に示す如

き単語頻度テーブル407に格納する。ただし、第15図に示す如きストップワードファイル7で指定されている単語（これをストップワードという）の出現頻度は求めない。なお、ストップワードファイル7におけるストップワードの指定は、前置詞、冠詞といった品詞での指定と、be, haveといった単語での指定の2つの指定方法で行なわれ、前者で指定したときは指定方法ラベルに“0”を設定し、後者で指定したときは指定方法ラベルに“1”を設定し、両者を区別する。

ステップ42では、予め設定された値以上の出現頻度を持つ単語を重要語として選ぶ。また、前記第1図のステップ11で文書の検索に検索キーが用いられたときは、その検索キーも重要語に加える。

ステップ43では、各文における重要語数を各文における単語数で除算して各文の得点を求める。求めた得点は、第16図に示す如き文得点テーブル408に格納する。

ステップ44では、得点の高い順に文を予め設

- 11 -

定した数だけ取り出し、それらの文（これらを抄録文という）を出現順に並べて抄録とする。実際には、抄録テーブルを作成する。

抄録の作成方法の他の例としては、例えば文書中の各段落の先頭文を取り出し順に並べて抄録とする方法を用いることが出来る。この方法は最も簡単であるため、簡便な抄録で足る場合に有用である。また、例えば「運用Ⅱ・人文研究のための言語データ処理入門／朝倉日本語新講座6 pp. 2-4（朝倉書店刊、1983）」に開示の方法を利用できる。さらに、特開昭61-117658公報に開示の方法を利用できる。

第1図に戻り、ステップ45では、各抄録文に順に抄録文番号を付し、その抄録文番号と元の文番号とを対応づけて、第17図に示す如き抄録文テーブル409を作成する。

ステップ14では、各抄録文に対応する元の文書中の文章の範囲を決定する。

この対応範囲決定処理を第3図を参照して説明する。

- 12 -

ステップ91では、抄録文番号 i (A_i) を“1” (A_1) に初期設定する。

ステップ92では、抄録文番号 i (A_i) に対応する元の文番号 j (B_j) を第17図の抄録文テーブル409から求める。 $i = 1$ (A_1) なら、 $j = 2$ (B_2) が求まる。

ステップ93では、文番号 j (B_j) の文が属する段落の段落番号 k (D_k) を第12図の管理情報テーブル61から求める。 $j = 2$ (B_2) なら、 $k = 1$ (D_1) が求まる。

ステップ94では、段落番号 k (D_k) の段落の最初の文の文番号を前記管理情報テーブル61から求めて、抄録文番号 i (A_i) の抄録文に対応する元の文書中の文章の範囲の開始位置とする。

ステップ95では、段落番号 k (D_k) の段落の最後の文の文番号を前記管理情報テーブル61から求めて、抄録文番号 i (A_i) の抄録文に対応する元の文書中の文章の範囲の終了位置とする。

ステップ96では、前記抄録文番号 i (A_i) の抄録文に対応する文章の範囲を拡張する。

- 13 -

- 590 -

- 14 -

この対応範囲拡張処理を第4図を参照して説明する。

ステップ111では、前記開始位置の文が、接続語句ファイル8に登録されている接続語句を含んでいるかどうか調べる。接続語句を含んでおればステップ112に進み、含んでいなければステップ114に進む。

接続語句ファイル8には、第18図に示すように、接続語句（接続詞やそれに準ずる語句）とその接続語句の性質に応じた対応文章範囲の拡張幅とが格納されている。例えば接続語句“but”とそれに応じた拡張幅“-1”とが格納されている。これは接続語句“but”を含む文が開始位置の文なら、その前の文から読む必要があることを意味している。

ステップ112では、拡張幅が負か調べる。負ならステップ113に進み、負でないならステップ114に進む。

ステップ113では、（開始位置+拡張幅）を新たな開始位置とする。例えば開始位置の文が接

続語句“but”を含んでいたなら、開始位置が1文前に拡張される。

ステップ114では、前記終了位置の文が、接続語句ファイル8に登録されている接続語句を含んでいるかどうか調べる。接続語句を含んでおればステップ115に進み、含んでいなければ対応範囲拡張処理を終了する。

ステップ115では、拡張幅が正か調べる。正ならステップ116に進み、正でないなら対応範囲拡張処理を終了する。

ステップ116では、（終了位置+拡張幅）を新たな終了位置とする。例えば終了位置の文が接続語句“as follows”を含んでいたなら、終了位置が1文後に拡張される。そして、対応範囲拡張処理を終了する。

第3図に戻り、ステップ97では、抄録文とその対応範囲を第19図に示す如き対応範囲テーブル410に登録する。

ステップ98では、抄録文番号iをインクリメントする。

- 15 -

ステップ99では、抄録文番号iが抄録文テーブル409の最後を過ぎたか調べる。過ぎていなければ前記ステップ92に戻り、過ぎていれば対応範囲決定処理を終了する。

第1図に戻り、ステップ15では、抄録を表示する。表示例を第22図に示す。

ステップ16では、テキスト情報を表示することをユーザが選択したかチェックする。選択したならステップ17に進み、選択しなければ処理を終了する。

ステップ17では、テキスト情報を抽出し、表示する。

このテキスト情報抽出・表示処理を第5図を参照して説明する。

ステップ151では、表示された抄録の中からテキスト情報を得たい抄録文をユーザが選択する。

ステップ152では、選択された抄録文とその前後のいくつかの抄録文に対応する文章の範囲を第19図の対応範囲テーブル410から求めて、それぞれ処理範囲とする。

- 16 -

ステップ153では、処理モードをユーザが選択する。処理モードには、原文表示処理と、キーワード検索処理と、他文書検索処理と、抄録文選択処理と、終了とがあり、これらのいずれかを選択すると、ステップ154～ステップ157のいずれかに進むか又はテキスト情報抽出・表示処理を終了する。

ステップ154では、原文表示処理を行なう。この原文表示処理を第6図を参照して説明する。

ステップ161では、抄録文と前記処理範囲の文章を対照表示するか否かをユーザに選択させる。対照表示を選択するとステップ162に進み、選択しなければステップ165に進む。

ステップ162では、画面を左右に分割する。

ステップ163では、前記第5図のステップ151でユーザが選択した抄録文を含む抄録の部分画面の左側エリアに表示する。この際、ユーザが選択した抄録文が左側エリアの中心の行に位置するように、また、輝度を強くするなどにより他と識別可能に表示する。さらに、前記左側エリア

に表示した抄録文に対応する処理範囲の文章を英文テキストファイル6の文書60から読み込んで、画面の右側エリアに表示する。前記ユーザが選択した抄録文に対応する文章中のユーザが選択した文は、表示輝度を強くするなどにより他と識別可能に、また、ユーザが選択した文が右側のエリアの中心の行に位置するように表示する。このときの画面の一例を第23図に示す。

一方、ステップ165では、前記処理範囲の文章を英文テキストファイル6の文書60から読み込んで、画面に表示する。この際、ユーザが選択した抄録文に対応する文章中の文が画面の中心の行に位置するように、また、輝度を強くするなどにより他と識別可能に表示する。このときの画面の一例を第24図に示す。

ステップ167では、ユーザが選択した抄録文に対応する文章が適正か否かをユーザが判断し、適正なら処理範囲を変更しないと入力し、適正でないなら処理範囲を変更すると入力する。処理範囲を変更しないと入力されたら原文表示処理を終

了する。処理範囲を変更すると入力されたらステップ168に進む。

ステップ168では、ユーザの指示に基づいて処理範囲の開始位置と終了位置を変更する。処理範囲の変更は、第19図の対応範囲テーブル410の対応範囲の開始位置と終了位置をユーザの指示に従って書き換えることによって行う。そして、前記ステップ161に戻る。

第5図に戻り、ステップ155では、キーワード検索処理を行なう。このキーワード検索処理を第7図を参照して説明する。

ステップ191では、キーワードを入力するための参考とする単語リストを表示するか否かをユーザに選択させる。表示するならばステップ192に進み、表示しないならばステップ193に進む。

ステップ192では、単語リストを表示する。この単語リスト表示処理を第8図を参照して説明する。

ステップ201では、表示する単語リストの種

類をユーザに選択させる。表示する単語リストの種類には、出現頻度順単語リストおよびアルファベット順単語リストの2種類がある。

ステップ202では、前記処理範囲の文章に出現する単語を第13図の出現単語テーブル406から得て、第20図の表示単語テーブル411に格納する。但し、ストップワードファイル7で指定されている単語は除く。また、この時点では、表示単語テーブル411の各単語の出現頻度の欄は、初期値として“1”を設定しておく。

ステップ203では、表示単語テーブル411中に重複して出現する単語を探し、その数をカウントし、その単語を含むレコードを一つだけ残して他を削除し、残したレコードの出現頻度を前記カウントした数で置き換える。

ステップ204では、前記ステップ201で出現頻度順単語リストが指定されたか否かを調べる。指定されている場合にはステップ205に進み、指定されていない場合にはステップ206に進む。

ステップ205では、表示単語テーブル411

を出現頻度順にソートする。

一方、ステップ206では、表示単語テーブル411をアルファベット順にソートする。

ステップ207では、表示単語テーブル411をリスト化して表示する。第25(a)図に出現頻度順単語リストを例示する。また、第25(b)図にアルファベット順単語リストを例示する。

第7図に戻り、ステップ193では、キーワードをユーザが入力する。この入力、前記ステップ192で表示された単語リスト中の単語を選択することにより入力してもよいし、キーボード装置3から入力してもよい。また、複数のキーワードを入力してもよい。キーワードが複数ある場合には、処理が繰返しになるだけである。入力されたキーワードは、第21図に示す如きキーワードテーブル412に格納する。

ステップ194では、前記キーワードテーブル412に格納したキーワードを含む文を処理範囲内の文章から抽出する。キーワードを含む文の文番号は、第13図の出現単語テーブル406を調

べることによって得られる。

ステップ195では、前記抽出したキーワードを含む文を英文テキストファイル6の文書60から取り出して、表示する。表示方法としては、第26(a)図に示すように、キーワードを含む文のみを表示してもよいし、第26(b)図に示すように、処理範囲の文章を表示しておき、キーワードを含む文をリンク等により強調してもよい。

ステップ196では、第5図のステップ151でユーザが選択した抄録文に対応する処理範囲が適正か否かをユーザが判断し、適正なら処理範囲を変更しないと入力し、適正でないなら処理範囲を変更すると入力する。処理範囲を変更しないと入力されたらキーワード検索処理を終了する。処理範囲を変更すると入力されたらステップ197に進む。

ステップ197では、ユーザの指示に基づいて処理範囲の開始位置と終了位置を変更する。処理範囲を変更すると、それにより第19図の対応範囲テーブル410の対応範囲の開始位置と終了位

置が更新される。そして、前記ステップ194に戻る。

第5図に戻り、ステップ156では、他文書検索処理を行なう。この他文書検索処理を第9図を参照して説明する。

ステップ251では、キーワードを入力する参考とするために単語リストを表示するか否かをユーザに選択させる。表示するのならば、ステップ252に進む。表示しないのならば、ステップ253に進む。このステップ251は、前記第7図のステップ191と同様である。

ステップ252では、単語リストを表示する。このステップ252は、前記第7図のステップ192と同様であり、具体的には第8図の処理となる。

ステップ253では、キーワードを入力する。このステップ253は、前記第7図のステップ193と同様である。

この後、第5図に戻ってから第1図のステップ11へ移行する。そして、キーワードテーブル4

— 23 —

12中の単語を検索キーとして他の文書を検索し、更にステップ12からステップ17の一連の処理を再帰的に行う。

この他文書検索処理によって、ある文書のテキスト情報を抽出している際に、関連する他の文書を連想的に検索できる。すなわち、ある文書中の単語や、ある文書を読んでいて思い付いた単語を検索キーとして、他の文書を連想的に検索できる。

さて、第5図に戻り、ステップ157の抄録文選択処理では、抄録を表示し、前記ステップ151に戻る。

以上の実施例では英文テキストファイル6の管理情報テーブル61に段落の情報を持っていたが、次に変形実施例として段落の情報を持たない場合を説明する。

この変形実施例の場合、第1図のステップ14の処理として、前記第3図のステップ91～ステップ99を行なう代わりに、第10図のステップ261～ステップ269を行なう。

すなわち、ステップ261では、ある抄録文に

対応する元の文書60の文を基準とした相対的な値で、その抄録文に対応する元の文書60の文章の範囲をユーザが設定する。具体的には、ある抄録文に対応する元の文書60の文を“0”とし、その文より前の文は負の値、その文より後の文は正の値で文の数をカウントして、対応範囲の相対的開始位置 α と相対的終了位置 β とを設定する。この設定は、予め行なわれてもよい。

ステップ262では、抄録文番号 i (A_i)を“1”に初期化する。

ステップ263では、抄録文番号 i (A_i)の抄録文の元の文番号 j (B_j)を第17図の抄録文テーブル409から求める。

ステップ264では、文番号($j + \alpha$)の文を、抄録文番号 i (A_i)の抄録文の対応範囲の開始位置とする。

ステップ265では、文番号($j + \beta$)の文を、抄録文番号 i (A_i)の抄録文の対応範囲の終了位置とする。

ステップ266では、第18図の接続語句ファ

— 24 —

イル8を参照して対応範囲を拡張する。この処理は第3図のステップ96と同様であり、具体的には第4図に示す処理となる。

ステップ267では、抄録文とその対応範囲を第19図に示す如き対応範囲テーブル410に登録する。

ステップ268では、抄録文番号iをインクリメントする。

ステップ269では、抄録文番号iが抄録文テーブル409の最後を過ぎたか調べる。過ぎていなければ前記ステップ263に戻り、過ぎていれば対応範囲決定処理を終了する。

他の変形実施例としては、第1図のステップ11～ステップ45を、ユーザが抄録を作成して入力するステップと、ユーザが第17図の抄録文テーブル409を作成して入力するステップとに置換したものが挙げられる。

上記実施例のテキスト情報抽出方法および装置によれば、抄録中で分かりにくかったり、特に関心が持たれた抄録文に関係する元の文章の一部を

的確に表示したり、出現単語リストを表示することにより、必要十分なテキスト情報を効率的に獲得することが出来る。

以上の第1実施例は抄録文に対応する元の文章を検索して表示するシステムであったが、次に第2実施例として抄録文とその翻訳文とを組み合わせで表示するシステムについて説明する。

第33図は、本発明の第2実施例のテキスト検索・表示システム2000のブロック図である。

図中、1は処理装置、2はディスプレイ装置、3はキーボード装置、4はメモリ、5は日本文テキストファイル、6は英文テキストファイル、7はストップワードファイル、8は接続語句ファイル、9は辞書ファイルである。

メモリ4は、文書検索プログラム401と、抄録作成プログラム402と、対応範囲決定プログラム403と、自動翻訳プログラム404と、テキスト情報抽出プログラム405と、出現単語テーブル406と、単語頻度テーブル407と、文得点テーブル408と、抄録文テーブル409と、

— 27 —

対応範囲テーブル410と、表示単語テーブル2411と、キーワードテーブル412とを含んでいる。

英文テキストファイル6は、複数の英文の文書60と、各文書毎の管理情報テーブル61とを含んでいる。この管理情報テーブル61は、第12図に示すように、文書の各文についての文番号とその文が属する段落の段落番号とを対応付けたものである。

第27図は、上記テキスト検索・表示システム2000において英文テキストファイル6から文書を検索し、抄録を作成し、その抄録を表示すると共に、その抄録を翻訳し、抄録と対応付けて表示する処理のフロー図である。

ステップ271では、英文テキストファイル6から文書を検索する。この処理は第1図のステップ11と同様である。

ステップ272では、前記取り出した文書の文を形態素解析し、各文の構成単語を第13図の出現単語テーブル406に出現順に格納する。この

処理は第1図のステップ12と同様である。

ステップ273では、前記文書の抄録を作成する。この処理は第1図のステップ13と同様である。

ステップ2745では、各抄録文に順に抄録文番号を付し、その抄録文番号と元の文番号とを対応づけて、第17図に示す如き抄録文テーブル409を作成する。この処理は第1図のステップ45と同様である。

ステップ274では、各抄録文に対応する元の文書中の文章の範囲を決定する。この処理は第1図のステップ14と同様である。

ステップ275では、抄録文に対し翻訳処理を行う。翻訳処理は、例えば特開昭58-40684号公報に開示の翻訳方法を利用することが出来る。翻訳文は、抄録文の元の文番号と対応付けて日本文テキストファイル5に格納する。英文テキストファイル6の文書60のデータ構造を第34(a)図に示し、日本文テキストファイル5のデータ構造を第34(b)図に示す。

— 28 —

ステップ276では、抄録を表示する。その際、ユーザは、表示言語選択キーにより抄録文を表示する言語として英語または日本語を選択する。英語が選択されたら、英文テキストファイル6の文書60から抄録文になっている文を抽出して表示する。この英語での表示例を第36(a)図に示す。日本語が選択されたら、日本語テキストファイル5から翻訳文を抽出して表示する。この日本語での表示例を第36(b)図に示す。

ステップ277では、抄録を読んだユーザが、テキスト情報を表示するか否かを入力する。この処理は、第1図のステップ16と同様である。表示するならばステップ278に進み、表示しないならば処理を終了する。

ステップ278では、テキスト情報を抽出し、表示する。このテキスト情報抽出・表示処理を第28図を参照して説明する。

ステップ301では、表示された抄録の中からテキスト情報を得たい抄録文をユーザが選択する。

ステップ302では、選択された抄録文とその

前後のいくつかの抄録文に対応する文章の範囲を第19図の対応範囲テーブル410から求めて、それぞれ処理範囲とする。

ステップ303では、処理範囲の文章に翻訳処理を行なう。翻訳文は、日本語テキストファイル5に格納する。また、翻訳処理で得られた英語出現単語と訳語を第35図に示す如き表示単語テーブル2411に格納する。この時点では、出現頻度の欄は初期値として“1”を設定しておく。

ステップ304では、処理モードをユーザが選択する。処理モードには、原文表示処理と、キーワード検索処理と、他文書検索処理と、抄録文選択処理と、終了とがあり、これらのいずれかを選択すると、ステップ305～ステップ308のいずれかに進むか又はテキスト情報抽出・表示処理を終了する。

ステップ305では、原文表示処理を行なう。この原文表示処理は、第6図を参照して先述した処理と同様である。但し、ユーザは、表示言語選択キーにより原文を表示する言語として英語また

— 31 —

は日本語を選択することが出来る。英語が選択されたら、英文テキストファイル6の文書60から処理範囲の文を抽出して表示する。この場合の処理範囲の文のみの表示例は第24図のようになる。また、抄録文と処理範囲の文の対照表示の例は第23図のようになる。一方、日本語が選択されたら、日本語テキストファイル5から前記処理範囲に対応する翻訳文を抽出して表示する。この場合の処理範囲に対応する翻訳文のみの表示例を第37(a)図に示す。また、抄録文の翻訳文と処理範囲の翻訳文の対照表示の例を第37(b)図に示す。

ステップ306では、キーワード検索処理を行なう。このキーワード検索処理を第29図を参照して説明する。

ステップ341では、テキスト情報を検索するためのキーワードを設定するための参考とする単語リストを表示するか否かをユーザに選択させる。表示するならばステップ342に進み、表示しないならばステップ343に進む。

ステップ342では、単語リストを表示する。

— 32 —

この単語リスト表示処理を第30図を参照して説明する。

ステップ3501では、表示する単語リストの種類をユーザに選択させる。表示する単語リストの種類は、出現頻度順単語リストおよびアルファベット/50音順単語リストの2種類がある。

ステップ3502では、前記第28図のステップ303で作成した第35図の表示単語テーブル2411中で英語出現単語と訳語の両方が一致する単語を探し、複数個ある場合にはその数をカウントし、そのレコードを1つ残して他を削除し、残したレコードの出現頻度の欄にカウントした数を設定する。

ステップ3503では、単語を表示する言語として英語または日本語を表示言語選択キーによりユーザに選択させる。英語が選択されたら、ステップ3504に進む。日本語が選択されたら、ステップ3508に進む。

ステップ3504では、前記ステップ3501で出現頻度順単語リストが選択されたか否かを調

— 33 —

— 595 —

— 34 —

べ、選択されている場合はステップ3505に進み、選択されていない場合はステップ3506に進む。

ステップ3505では、第35図の表示単語テーブル2411の英語出現単語を出現頻度順にソートする。

ステップ3506では、第35図の表示単語テーブル411の英語出現単語をアルファベット順にソートする。

ステップ3507では、ソートした表示単語テーブル2411の英語出現単語を表示する。第38(a)図に英語出現単語の出現頻度順単語リストを例示する。また、第38(b)図に英語出現単語のアルファベット順単語リストを例示する。

一方、ステップ3508では、前記ステップ3501で出現頻度順単語リストが選択されたか否かを調べ、選択されている場合はステップ3509に進み、選択されていない場合はステップ3510に進む。

ステップ3509では、第35図の表示単語テ

ーブル2411の訳語を出現頻度順にソートする。

ステップ3510では、第35図の表示単語テーブル411の訳語を50音順にソートする。

ステップ3511では、ソートした表示単語テーブル2411の訳語を表示する。第39(a)図に訳語の出現頻度順単語リストを例示する。また、第39(b)図に訳語の50音順単語リストを例示する。

第29図に戻り、ステップ343では、キーワードを設定する。このキーワード設定処理を第31図を参照して説明する。

ステップ381では、ユーザがキーワードを入力する。キーワードは、英語、日本語どちらでもよく、複数でもよい。

ステップ382では、ユーザが入力したキーワードが英語か否かを識別する。具体的には、キーワードがキーボード装置3から入力された場合にはキーボード装置3の入力モードから英語か否かを識別する。また、表示している単語リストから選択された場合には、その単語リストの種類から

英語か否かを識別する。キーワードが英語の場合にはステップ383に進み、日本語の場合にはステップ384に進む。

ステップ383では、英語のキーワードをキーワードテーブル412に格納する。

一方、ステップ384では、日本語のキーワードと同じ訳語を持つレコードを第35図の表示単語テーブル2411で探す。

ステップ385では、日本語のキーワードと同じ訳語を持つレコードが表示単語テーブル2411にあったかチェックする。あったならばステップ386に進む。なかったならば、前記ステップ381に戻り、キーワードを再入力させる。

ステップ386では、日本語のキーワードと同じ訳語を持つレコードの英語出現単語をキーワードテーブル412にキーワードとして格納する。もし、日本語のキーワードと同じ訳語を持つレコードが複数ある場合には、その全ての英語出現単語をキーワードテーブル412にキーワードとして格納する。

第29図に戻り、ステップ344では、前記キーワードテーブル412に格納したキーワードを含む文の文番号を処理範囲内の文章から抽出する。この処理は前記第7図のステップ194と同様である。

ステップ345では、前記抽出したキーワードを含む文を英文テキストファイル6の文書60から取り出して、表示する。この処理は前記第7図のステップ195と同様である。

ステップ346では、処理範囲を変更するか否かをユーザに入力させ、処理範囲を変更しないと入力されたらキーワード検索処理を終了する。処理範囲を変更すると入力されたらステップ347に進む。この処理は前記第7図のステップ196と同様である。

ステップ347では、ユーザの指示に基づいて処理範囲の開始位置と終了位置を変更する。処理範囲を変更すると、それにより第19図の対応範囲テーブル410の対応範囲の開始位置と終了位置が更新される。そして、前記ステップ344に

戻る。

さて、第28図に戻り、ステップ307では、他文書検索処理を行なう。この他文書検索処理を第32図を参照して説明する。

ステップ4251では、キーワードを設定する参考とするために単語リストを表示するか否かをユーザに選択させる。表示するのならば、ステップ4252に進む。表示しないのならば、ステップ4253に進む。このステップ4251は、前記第29図のステップ341と同様である。

ステップ4252では、単語リストを表示する。このステップ4252は、前記第29図のステップ342と同様であり、具体的には第30図の処理となる。

ステップ4253では、キーワードを設定する。このステップ4253は、前記第29図のステップ343と同様であり、具体的には第31図の処理となる。

この後、第28図に戻ってから第27図のステップ271へ移行する。そして、キーワードテ

ブル412中の単語を検索キーとして他の文書を検索し、更にステップ272からステップ278の一連の処理を再帰的に行う。

第28図に戻り、ステップ308では、抄録文選択処理を行なう。すなわち、抄録を表示し、前記ステップ301に戻る。

変形実施例としては、第27図のステップ271～ステップ2745を、ユーザが抄録を作成して入力するステップと、ユーザが第17図の抄録文テーブル409を作成して入力するステップとに置換したものが挙げられる。

他の実施例としては、英語や日本語以外の他の言語について本発明を適用したものが挙げられる。
[発明の効果]

本発明のテキスト情報抽出方法および装置によれば、ある文書における重要箇所とか、その重要箇所に出現する単語のリストとかのテキスト情報を、抄録を利用して容易に得られるようになる。

4. 図面の簡単な説明

第1図は本発明の第1実施例の処理の基本フロ

— 39 —

— 40 —

ー図、第2図は抄録作成処理のフロー図、第3図は対応範囲決定処理のフロー図、第4図は対応範囲拡張処理のフロー図、第5図はテキスト情報抽出・表示処理のフロー図、第6図は原文表示処理のフロー図、第7図はキーワード検索処理のフロー図、第8図は単語リスト表示処理のフロー図、第9図は他文書検索処理のフロー図、第10図は対応範囲決定処理の他の例のフロー図、第11図は本発明の第1実施例のハードウェア構成図、第12図は情報管理テーブルの概念図、第13図は出現単語テーブルの概念図、第14図は単語頻度テーブルの概念図、第15図はストップワードファイルの概念図、第16図は文得点テーブルの概念図、第17図は抄録文テーブルの概念図、第18図は接続語句ファイルの概念図、第19図は対応範囲テーブルの概念図、第20図は表示単語テーブルの概念図、第21図はキーワードテーブルの概念図、第22図、第23図、第24図、第25(a)図、第25(b)図、第26(a)図および第26(b)図は表示画面の例示図、第27図は本発

明の第2実施例の処理の基本フロー図、第28図はテキスト情報抽出・表示処理のフロー図、第29図はキーワード検索処理のフロー図、第30図は単語リスト表示処理のフロー図、第31図はキーワード設定処理のフロー図、第32図は他文書検索処理のフロー図、第33図は本発明の第2実施例のハードウェア構成図、第34(a)図は英文の文書の概念図、第34(b)図は日本文テキストファイルの概念図、第35図は表示単語テーブルの概念図、第36(a)図、第36(b)図、第37(a)図、第37(b)図、第38(a)図、第38(b)図、第39(a)図および第39(b)図は表示画面の例示図である。

(符号の説明)

1・・・処理装置、2・・・ディスプレイ装置、3・・・キーボード装置、4・・・メモリ、401・・・文書検索プログラム、402・・・抄録作成プログラム、403・・・対応範囲決定プログラム、404・・・自動翻訳プログラム、405・・・テキスト情報抽出プログラム、406・

— 41 —

— 597 —

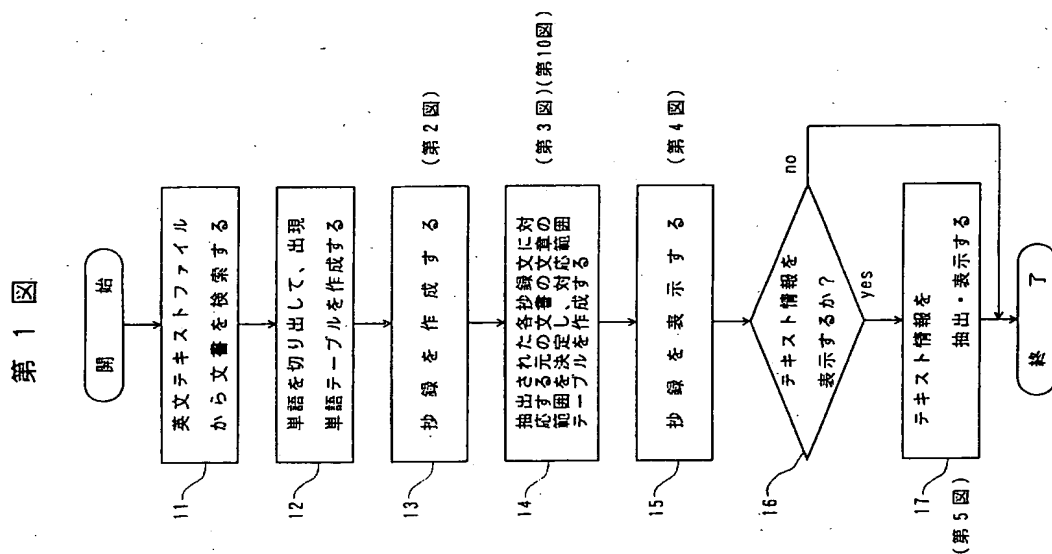
— 42 —

・・出現単語テーブル、407・・・単語頻度テーブル、408・・・文得点テーブル、409・
 ・・抄録文テーブル、410・・・対応範囲テーブル、411・・・表示単語テーブル、412・
 ・・キーワードテーブル、5・・・日本文テキストファイル、6・・・英文テキストファイル、6
 0・・・文書、61・・・管理情報テーブル、7
 ・・・ストップワードファイル、8・・・接続語
 句ファイル、9・・・辞書ファイル。

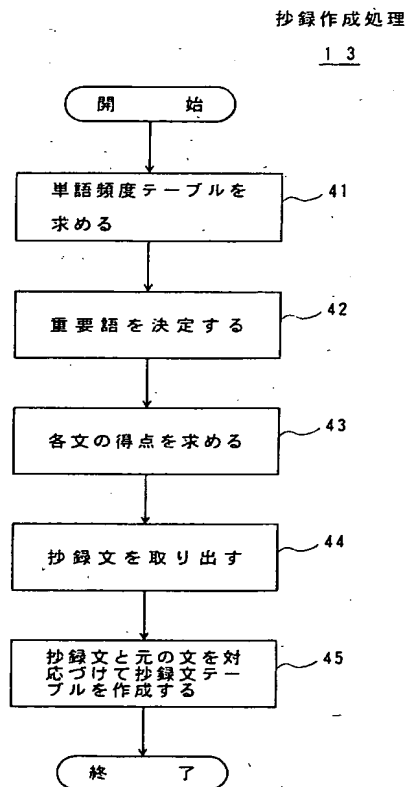
出願人 株式会社 日立製作所

代理人 弁理士 有近 紳志郎

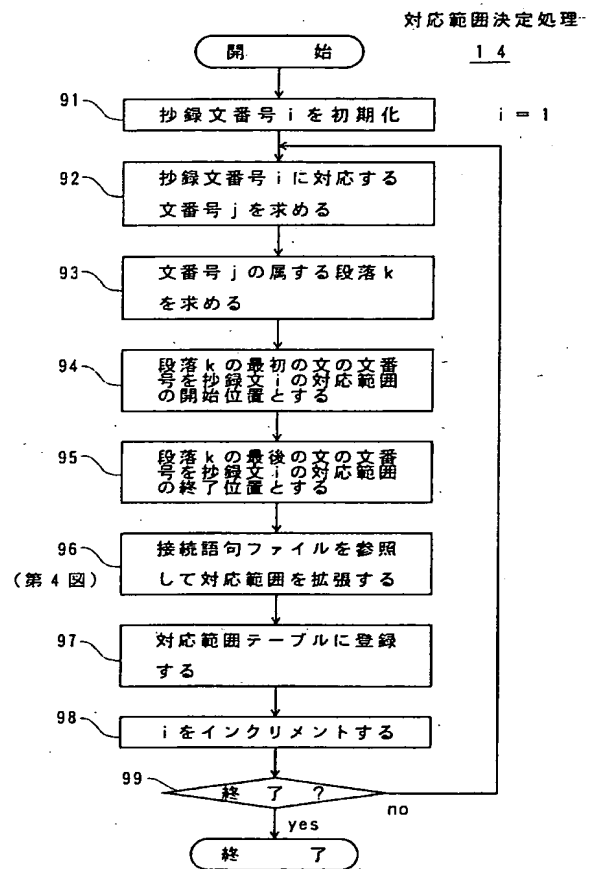
- 43 -



第 2 図

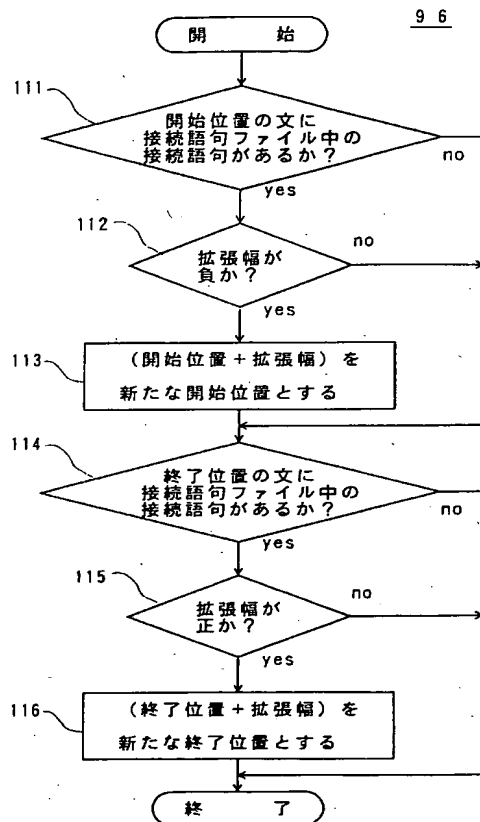


第 3 図



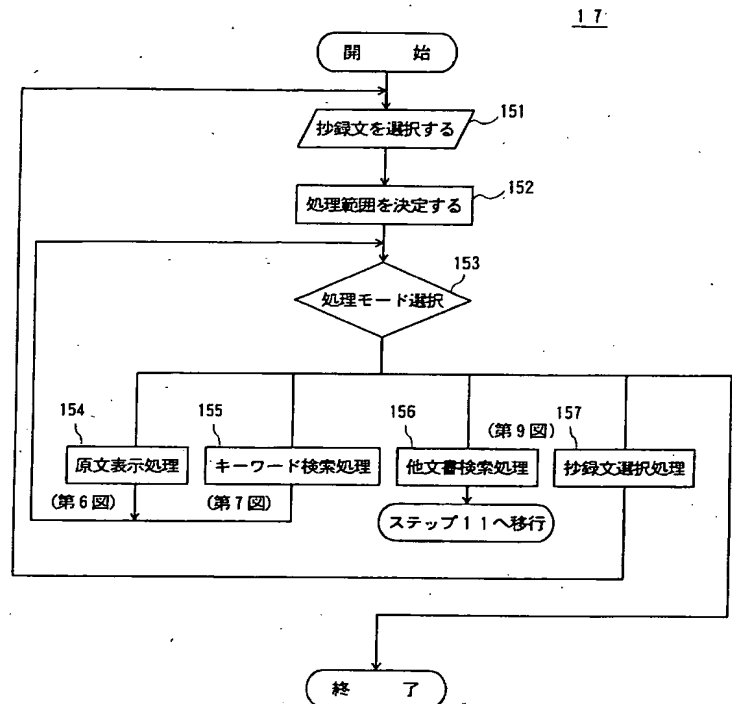
第 4 図

対応範囲拡張処理



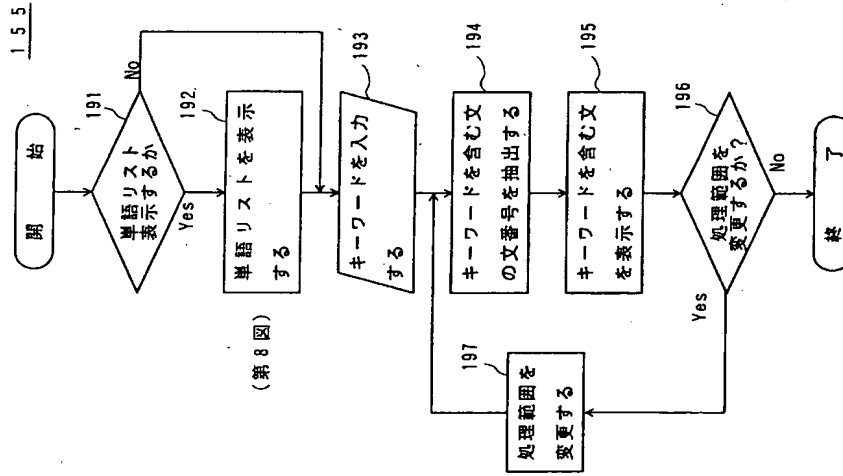
第 5 図

テキスト情報抽出・表示処理



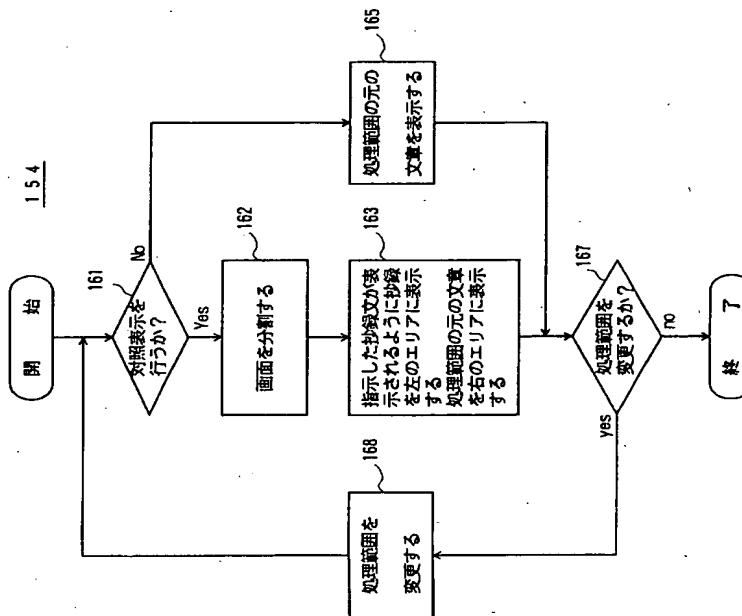
第 7 図

キーワード検索処理



第 6 図

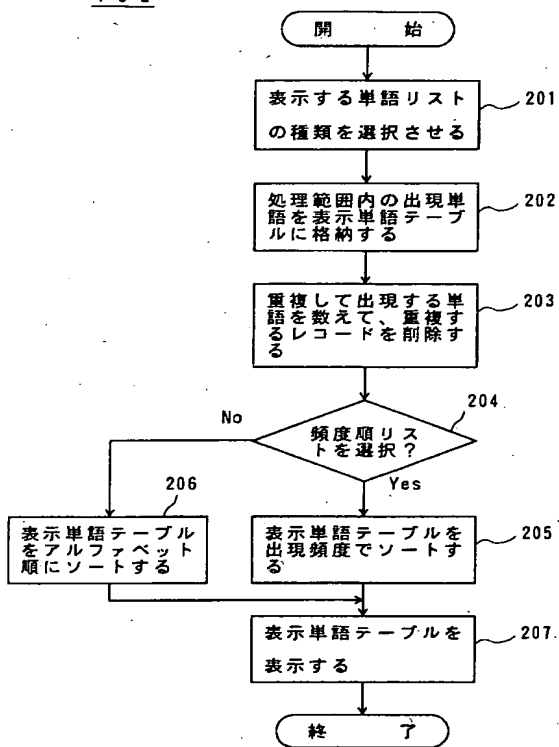
原文表示処理



第 8 図

単語リスト表示処理

192



第 9 図

他文書検索処理

156

(第20図)

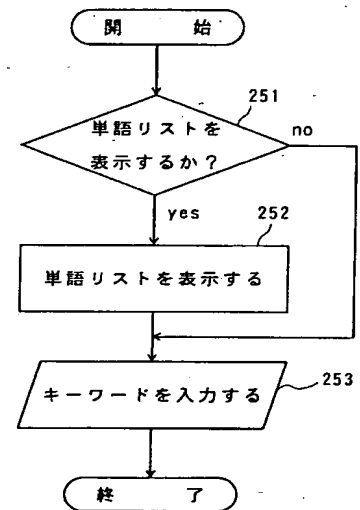
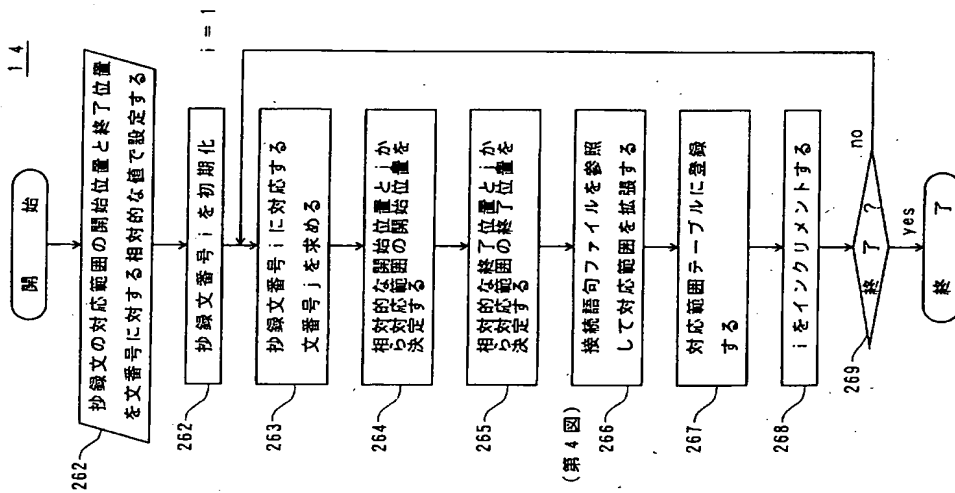


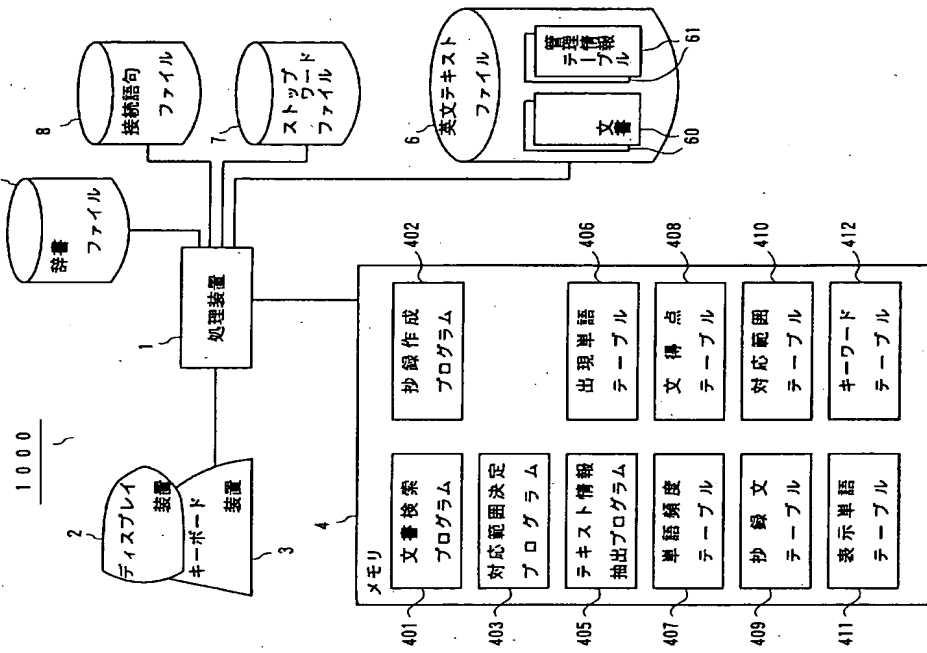
圖 10 紙

暨友誼編輯部



二、☐ 無

テキスト検索・表示システム



第 13 図

出現単語テーブル
4 0 6

文番号	出 現 単 語			
B 1	Japan	role	in	...
B 2	This	movement	have	...
B 3	The	number	of	...
B 4	it	be	very	...
⋮	⋮	⋮	⋮	⋮

文書
6 0

文番号	英 文 デ ー タ
B 1	Japan's role in the world has expanded dramatically.
B 2	This movement has necessitated translation of ...
B 3	The number of translators in the labor market ...
B 4	It is very ...

第 12 図

管理情報テーブル

6 1

文 番 号	段 落 番 号
B 1	D 1
B 2	D 1
B 3	D 1
B 4	D 2
⋮	⋮

第 15 図

ストップワードファイル

7

指定方法ラベル	ストップワード
0	前置詞
0	冠 詞
1	be
1	have
⋮	⋮

第 14 図

単語頻度テーブル

4 0 7

出 現 単 語	出 現 頻 度
translation	4
HICATS	3
document	2
system	2
⋮	⋮

第 16 図

文得点テーブル
408

文番号	重要語数	単語数	得点
B1	0	8	0
B2	2	11	0.182
B3	0	14	0
B4	1	14	0.071
⋮	⋮	⋮	⋮

第 18 図

接続語句ファイル
8

接続語句	拡張幅
but	-1
however	-1
as follows	-1
therefore	-1
⋮	⋮

第 20 図

表示単語テーブル
411

表示単語	出現頻度
HICATS	2
system	2
translation	2
⋮	⋮

第 17 図

抄録文テーブル
409

抄録文番号	文番号
A1	B2
A2	B5
A3	B7
⋮	⋮

第 19 図

対応範囲テーブル
410

抄録文番号	対応範囲開始位置	対応範囲終了位置
A1	B1	B3
A2	B4	B6
A3	B6	B7
⋮	⋮	⋮

第 21 図

キーワードテーブル
412

キーワード
HICATS
translation
cost
⋮

第 22 図

ディスプレイ装置

A1: This movement has necessitated translation of a large number of documents.

A2: Hitachi's Machine Translation System (HICATS) is a system aimed at just this problem.

A3: Moreover, HICATS can reduce the cost of translation, too.

第 23 図

<p>A1: This movement has necessitated ...</p> <p>A2: Hitachi's Machine Translation System...</p> <p>A3: Moreover, HICATS can reduce the cost...</p>	<p>#4: It is very difficult to translate...</p> <p>#5: Hitachi's Machine Translation System...</p> <p>#6: HICATS can increase the capacity of...</p>
---	--

第 24 図

#4: It is very difficult to translate all the documents that need to be translated.

#5: Hitachi's Machine Translation System (HICATS) is a system aimed at just this problem.

#6: HICATS can increase the capacity of translation.

第 25(a) 図

²

HICATS	system	translation
document	Hitachi	machine
problem	...	

第 26(a) 図

HICATS を含む文は

B5: Hitachi's Machine Translation System (HICATS)
is a system aimed at just this problem.

B6: HICATS can increase the capacity of translation.

第 25(b) 図

²

aim	capacity	cost
current	demand	difficult
document	...	

第 26(b) 図

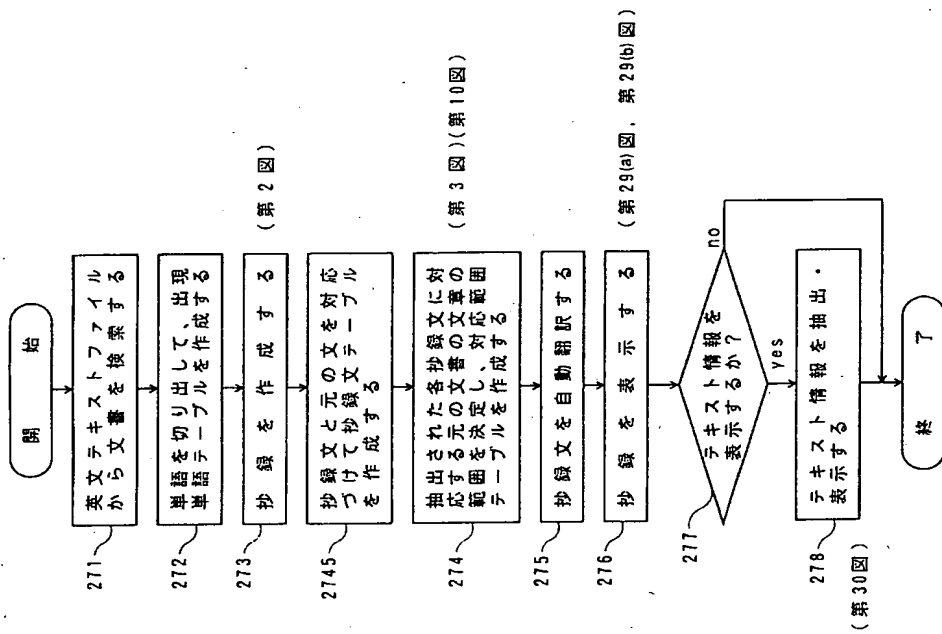
HICATS を含む文は

B4: It is very difficult to translate all the documents
that need to be translated.

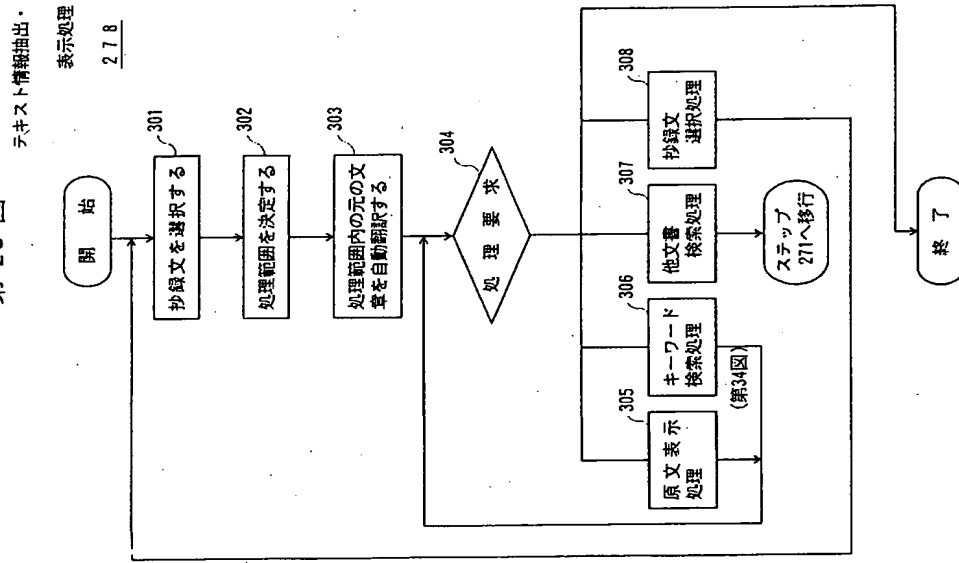
B5: Hitachi's Machine Translation System (HICATS)
is a system aimed at just this problem.

B6: HICATS can increase the capacity of translation.

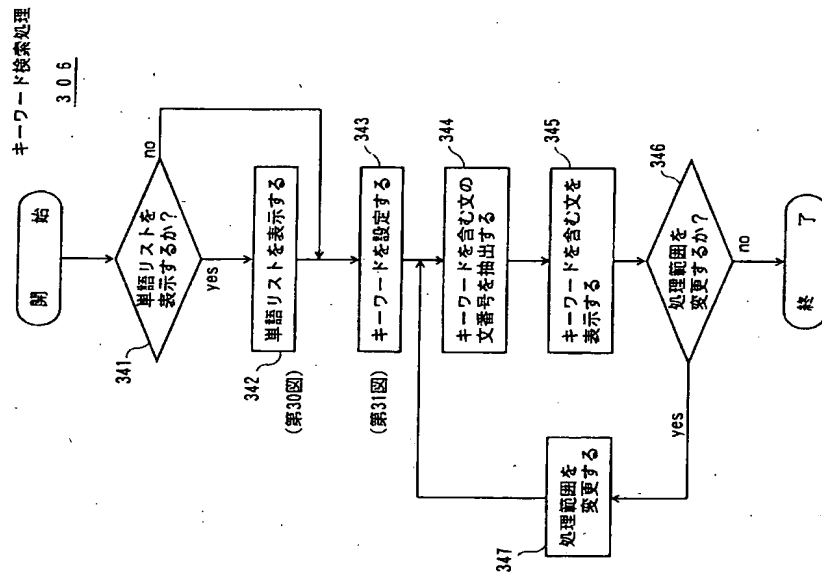
第 27 図



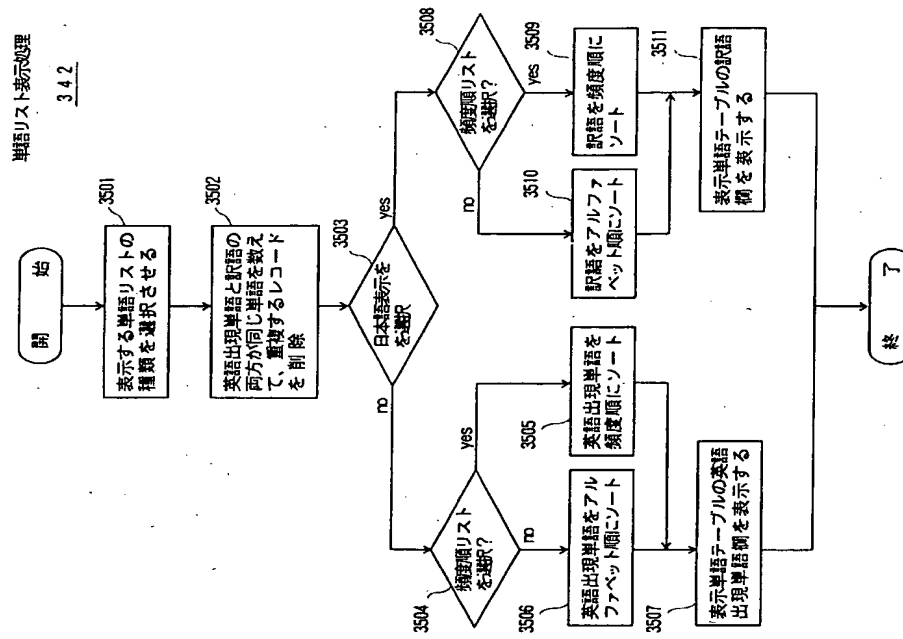
第 28 図



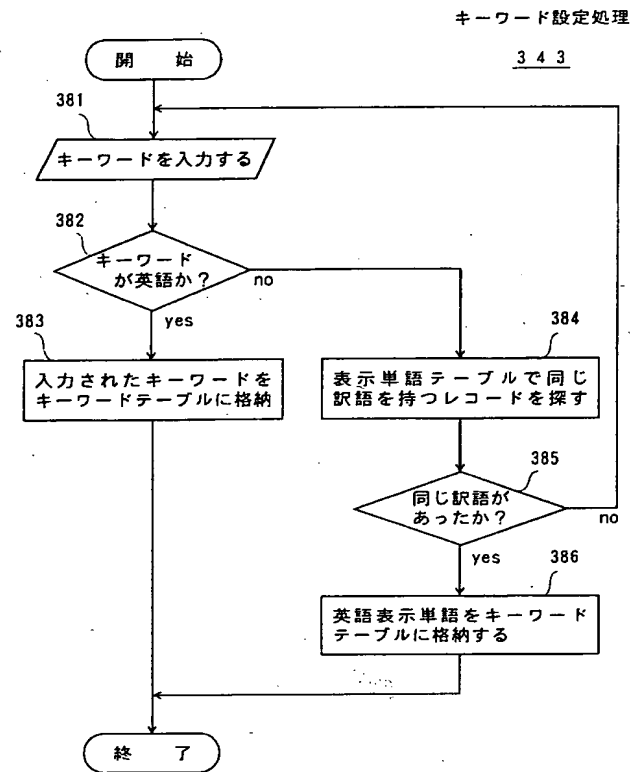
第 29 図



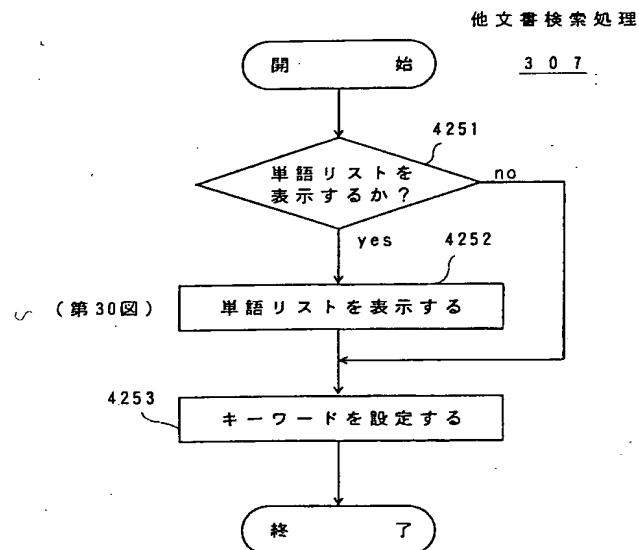
第 30 図



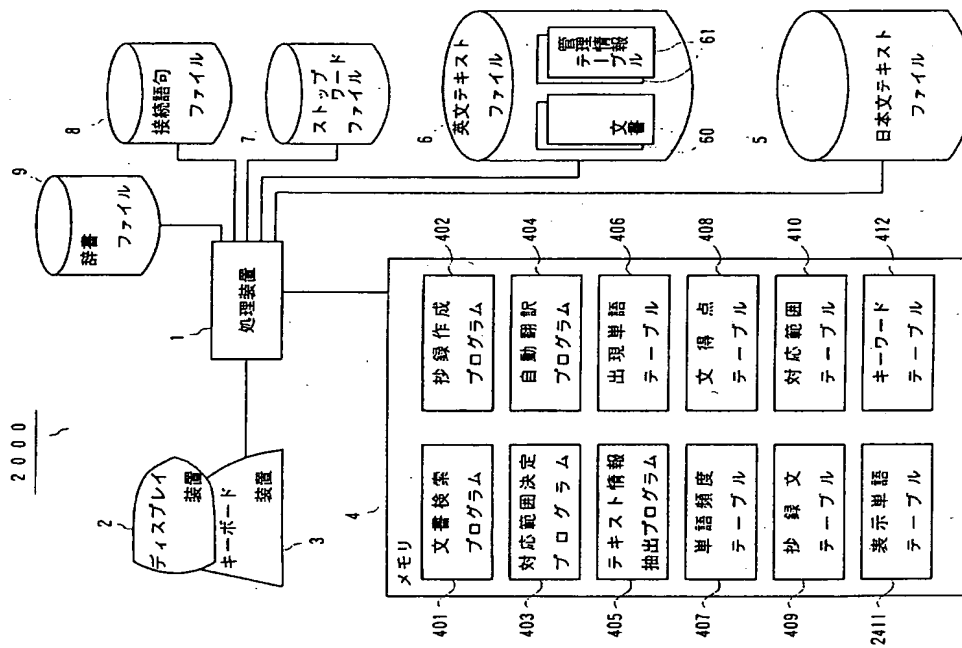
第 31 図



第 32 図



第 33 図



第 34(a)図

文書 60		日本語テキストファイル 5	
文番号	英文データ	文番号	日本語データ
B 1	Japan's role in ...	B 1	
B 2	This movement ...	B 2	この動きはたくさんの...
B 3	The number of ...	B 3	

第 35 図

表示単語テーブル 2411		
英語出現単語	訳 語	出現頻度
difficult	難しい	1
translate	翻訳する	2
document	ドキュメント	2
...

第 36(a) 図

A1: This movement has necessitated translation of a large number of documents.

A2: Hitachi's Machine Translation System (HICATS) is a system aimed at just this problem.

A3: Moreover, HICATS can reduce the cost of translation, too.

.

.

.

第 36(b) 図

A 1 : この動きはたくさんの文書の翻訳を必要とした。

A 2 : 日立の機械翻訳システム (HICATS) は、まさにこの問題に向けられるシステムである。

A 3 : その上、また、HICATSは翻訳の費用を下げることができる。

.

.

.

第 37(a) 図

4 : 翻訳される必要がある文書をすべて翻訳することは大変難しい。

5 : 日立の機械翻訳システム (HICATS) は、まさに、この問題に向けられるシステムである。

6 : HICATSは翻訳のキャパシティを増すことができる。

第 37(b) 図

²

A 1 : この動きはたくさんの文書の翻訳を必要とした。	# 4 : 翻訳される必要がある文書をすべて翻訳する…
A 2 : 日立の機械翻訳システム (HICATS) はまさにこの…	# 5 : 日立の機械翻訳システム (HICATS) はまさにこの…
A 3 : その上、また、HICATS は翻訳の費用を下げることに…	# 6 : HICATS は翻訳のキャパシティを増すことができる。

第 38(a) 図

²

HICATS	system	translation	document
Hitachi	machine	problem	...
		.	
		.	
		.	

第 38(b) 図

²

aim	capacity	cost	current
demand	difficult	document	...
		.	
		.	
		.	

第 39(a) 図

²

HICATS	システム	翻 訳	文 書
日立	機 械	問 題	...
	・		
	・		
	・		

第 39(b) 図

²

キャパシティ	現在の	コスト	ドキュメント
向ける	難しい	需 要	...
	・		
	・		
	・		